Polonca Kocjančič
Ljubljana

INTERNET Y LOS RECURSOS LINGÜÍSTICOS PARA LA LENGUA ESPAÑOLA: DICCIONARIOS Y CORPUS

Palabras clave: recursos lingüísticos, diccionario, corpus de referencia, Internet, español

1. Introducción

Las obras de referencia y los diversos recursos lingüísticos conforman un nexo de unión entre una lengua en su totalidad y el usuario que busca información sobre cualquier aspecto de la lengua que emplea, sea materna o extranjera. Las nuevas tecnologías han abierto un universo nuevo y representan un elemento que hay que considerar en la difusión de la información lingüística, con algunos recursos bien conocidos y otros nuevos, que investigan enfoques alternativos de la organización de la información lingüística.

Por *recursos lingüísticos* entendemos «léxicos, gramáticas y colecciones de textos o corpus» (Lavid, 2005: 61). En este artículo queremos presentar, ante todo, los recursos léxicos y léxico-gramaticales, por esa razón no presentaremos las gramáticas en el sentido de obras de referencia, aunque tenemos en cuenta que el sistema gramatical de una lengua forma parte indivisible de todos los recursos léxicos.

El *léxico* se entiende como un repositorio de palabras y de información sobre ellas, tales como sus características fonéticas (nivel fonológico), su estructura (nivel morfológico), sus posibilidades de combinación con otras palabras (nivel sintáctico), y su significado en diferentes contextos (nivel semántico). El almacén tradicional del conocimiento léxico sobre las palabras es el *diccionario*, cuyas posibilidades se ven ampliadas en la actualidad con la paralela creación de bases de datos léxicos en formato electrónico.

Un *corpus* es «una muestra amplia de lengua escrita o hablada que se considera representativa bien del estándar o de alguna variante diatípica o diatópica, o de algún período histórico determinado» (Lavid, 2005: 62). Entre los usos más frecuentes de corpus para un usuario humano destacan: el análisis de la lengua, la determinación de las características de la lengua analizada y la verificación empírica de teorías lingüísticas. Por otro lado, el corpus sirve al desarrollo de las *tecnologías de lenguaje humano* (TLH) o las *industrias de la lengua* y a la elaboración de nuevos recursos lingüísticos. El uso más frecuente en el campo del desarrollo de productos o servicios basados en el tratamiento del lenguaje es el entrenamiento de máquinas para adaptar su comportamiento a circunstancias específicas. Además, es posible utilizar los corpus como campo de pruebas de una aplicación de tecnología lingüística para poder determinar su buen funcionamiento en la práctica.

En el presente artículo nos centraremos en la enmarcación y descripción de una selección de los recursos que hoy en día existen en el campo de la lengua española para un público generalizado –aquél que busca información sobre esta lengua para cualquiera de las actividades lingüísticas–, con una ligera orientación hacia estudiantes de ELE y –futuros– traductores. El centro de nuestra atención lo constituirán los recursos que están a disposición de toda la comunidad lingüística, de acceso libre (y en la mayoría de los casos, gratuito o por un precio razonable) en Internet: diccionarios monolingües, diccionarios bilingües y corpus de referencia. Existen también varias bases de datos léxicos que también se ocupan de la lengua general, pero por razones de espacio no podemos tratarlas aquí (Spanish FrameNet, BDS, ADESSE, WordNet, AnCora etc.).

El artículo quiere invitar a los lectores a la investigación de varios enfoques, y posibilidades de presentación y visualización de la información lingüística; en ocasiones, es necesario familiarizarse primero con los recursos para obtener un resultado satisfactorio, no dejar de intentarlo si después de teclear una palabra clave uno no obtiene enseguida el resultado imaginado. Todos estos recursos sirven para efectuar paulatinamente un proceso de adquisición de conocimiento sobre el lenguaje.

2. Diccionarios

Los diccionarios no son libros de lectura, sino obras de consulta rápida que se utilizan para un fin concreto. Esta función suya exige que estén al servicio de quienes los consultan, y se presenten de tal forma que los usuarios accedan con la mayor rapidez y eficacia posible al significado que buscan (Almela et al., 2005). Los diccionarios en formato electrónico no son una excepción. Los lectores ya conocerán varios de los diccionarios presentes en línea, pero cabe enumerar algunos para los que no estén tan familiarizados con el tema, con la invitación a que investiguen los enlaces [para todos: fecha de consulta: 15 de junio de 2009]:

- El diccionario CLAVE de la editorial SM: http://clave.librosvivos.net/
- El diccionario Salamanca de la Lengua Española (DESAL, Santillana ELE): http://fenix.cnice.mec.es/diccionario/
- Diccionario de la lengua española de la RAE, 22ª edición con actualizaciones (DRAE): http://buscon.rae.es/draeI/
- Diccionario panhispánico de dudas, de la RAE (DPD): http://buscon.rae.es/dpdI/
- Nuevo Tesoro Lexicográfico de la Lengua Española (NTLLE): http://buscon.rae.es/ntlle/SrvltGUILoginNtlle (diccionarios de la RAE desde 1726 a 1992)
- Diccionarios Collins, entre otros el español-inglés, inglés-español: http://dictionary.reverso.net/
- Diccionarios WordReference: http://www.wordreference.com/ (Diccionario de la lengua española Espasa Calpe 2005, Pocket Oxford Spanish Dictionary 2005, Diccionario Espasa Concise: inglés-español 2000, Diccionario Espasa Grand: español-francés, français-espagnol 2000, Gran diccionario español-portugués português-espanhol 2001 etc.)

- Diccionarios VOX – portal de la empresa Larousse Editorial: http://www.diccionarios.com/ (Diccionario de Uso del Español de América (DUEAE 2005), Diccionario Sinónimos y Antónimos (2007), Diccionario Ideológico de la Lengua Española (2005) y los diccionarios bilingües entre el español y las siguientes lenguas: inglés, francés, alemán, catalán, italiano, portugués, gallego, eusquera; conjugador verbal para el español y el inglés; diccionario y tesauro inglés de la editorial Chambers; 25 consultas gratis, luego suscripción por 3, 6 o 12 meses)
- Diccionarios Collins, entre otros el español-inglés, inglés-español, versiones abreviadas: http://www.collinslanguage.com/shop/spanish-dictionary-landing.aspx
- Diccionarios ElMundo: http://www.elmundo.es/diccionarios/ (diccionario español-inglés, inglés-español, español-francés, francés-español, diccionario de la lengua española, sinónimos y antónimos)
- Diccionarios en http://www.diclib.com (entre otros, el diccionario de María Moliner)
- Diccionario PONS español-inglés, inglés-español, español-alemán, alemán-español http://www.pons.de/
- Diccionario Chambers Harrap y Velázquez Spanish and English Dictionary: http://www.spanishdict.com
- Diccionarios Merriam-Webster, inglés-español, español-inglés etc.: http://www.merriam-webster.com/
- Diccionario del español usual en México (DEUM): http://www.cervantesvirtual.com/servlet/SirveObras/35716130101359941976613/index.htm

Entre los diccionarios bilingües, los que mayor cobertura tienen son los diccionarios en http://dictionary.reverso.net/ y http://www.wordreference.com/, mientras que los demás analizados tienen menos información. En el Cuadro 1, véase la comparación del lema *duda* en varios diccionarios (incluso el *Google Dictionary* (http://www.google.com/dictionary):

DICCIONARIOS / FRASES	VOX español|inglés	PONS español|alemán	PONS español|inglés	ELMUNDO español|inglés	Spanishdict.com HARRAP	SpanishDict.comVELÁZQUEZ	Merriam Webster Spanish	Wordreference español|inglés	Collins Reverso español|inglés	Collinslanguage.com	Google Dictionary
beneficio de la duda											x
fuera de toda duda						x		x	x		
¡la duda ofende!					x				x		

no cabe la menor duda / no cabe duda		x	x		x	x	x	x	x	x	
no hay duda	x										x
no te (...) quepa duda	x				x				x	x	
poner algo en duda	x	x	x		x			x	x	x	x
sacar a algn de dudas / sacar de la duda	x				x			x	x	x	
salir de dudas	x	x	x		x			x	x		
sin (lugar a) duda				x	x			x	x		
sin duda (alguna)	x	x	x		x	x		x	x	x	x
sin la menor duda	x										x
tengo mis dudas					x			x	x	x	
vista de duda											x
NÚMERO DE EJEMPLOS ADICIONALES		2		4		1		7	**24**		

Cuadro 1: *La distribución de información lingüística en algunos diccionarios bilingües disponibles en Internet*

Hay también varios diccionarios con información lingüística especializada, por ejemplo:

- Diccionario de partículas discursivas del español (DDPD, de investigadores de varias universidades españolas): http://textodigital.com/P/DDPD/
- Diccionario de neologismos on line (Universidad Pompeu Fabra, encargado por la editorial LAROUSSE (DNOL; más de 4.000 lemas): http://obneo.iula.upf.edu/spes/

En algunos casos se trata de versiones informatizadas de obras anteriormente en formato de libro (HarperCollins, WordReference, VOX etc.); en otros casos se realizan actualizaciones directamente en la versión electrónica (por ejemplo el DRAE, el DESAL); y también hay algunos que sólo existen en línea (por ejemplo el DDPD). Cada obra sigue una política editorial determinada, que rige los criterios de inclusión/exclusión de material lingüístico. Así, muchas veces estos diccionarios sufren algunas de las desventajas de sus antecedentes en papel: por ejemplo, la falta de espacio. Este aspecto puede ser solventado en el caso de publicaciones elaboradas originariamente al formato electrónico. Hoy es muy fácil acceder a la información – además, todavía existe una amplia gama de diccionarios en papel o en cederrón que aquí no vamos a tratar. Las interpretaciones que proporcionan los diccionarios siempre tienen su valor. Pero, veamos lo que hay más allá del simple teclear una palabra en un diccionario en línea.

3. Corpus

El análisis de la lengua y el proceso de compilación de una obra de referencia siempre han requerido grandes cantidades de datos lingüísticos. A lo largo de la historia, ha sido enorme la tarea de construir manualmente una base organizada de datos sobre las palabras y sus contextos. Hoy en día, esta primera fase en la compilación de una obra de referencia ha sido facilitada en muchos aspectos por la informática. En las últimas décadas del siglo pasado, se han investigado y desarrollado procesos de compilación y análisis estadístico de datos lingüísticos; como consecuencia, se han construido muchos corpus para idiomas más diversos. Algunas ventajas de corpus son la anotación, la lematización y la representatividad, junto con las funciones de los programas de concordancias que permiten hacer diversos análisis estadísticos.

A continuación, presentaremos algunos de los corpus principales que existen para la lengua española en la actualidad, y luego pasaremos a describir las actuales líneas de desarrollo.

3.1 El Banco de datos de la Real Academia Española: el CREA - Corpus de Referencia del Español Actual

El CREA (*Corpus de Referencia del Español Actual*, http://corpus.rae.es/creanet.html) es un corpus representativo del estado actual de la lengua española elaborado por la RAE. El corpus se construyó a principios de los años noventa y se puso en marcha en 1995, seguido pocos meses después por el CORDE (*Corpus diacrónico del español*). Los dos corpus son complementarios: el CREA contiene textos entre 1975 y 2004, y el CORDE de los períodos anteriores a 1975. Periódicamente, al CREA se le añaden nuevos textos – la última serie fueron los textos del período 2000-2004, incorporados en junio de 2008, y está previsto que los textos del CREA pasen gradualmente al CORDE. Desde el pasado junio de 2008, el CREA contiene 154 279 050 formas de fuentes escritas y habladas de todos los países de habla española (50 % España, 50 % América Latina). A todos los materiales procesados, tanto en el CREA como en el CORDE, se les ha añadido una serie de marcas textuales, establecidas según el estándar internacional SGML (*Standard General Markup Language*), de acuerdo con las recomendaciones de la TEI (*Text Encoding Initiative*).

La parte oral del corpus contiene 9 millones de palabras y es accesible desde 2004. Reúne textos de colecciones anteriores (ACUAH, ALFAL, Caracas-77, Caracas-87, CEAP, COVJA, CSC, CSMV, UAM) y materiales nuevos obtenidos desde la radio, la televisión, el teléfono, y con grabaciones directas.

La interfaz de consulta de los dos corpus académicos se presenta abajo:

Cuadro 2: *Página principal de consulta en el CREA: «obtener dist/2 resultado»; en libros y revistas de España y los EE.UU*.

En la pantalla de consulta, existen los siguientes criterios de selección: Autor, Obra, Cronológico, Medio (libros, periódicos, revistas, miscelánea, oral), Geográfico (países de habla española), Tema (ciencias y tecnologías, ciencias sociales, creencias y pensamiento, política, economía, comercio, finanzas, artes, ocio, vida cotidiana, salud, ficción, miscelánea, oral).

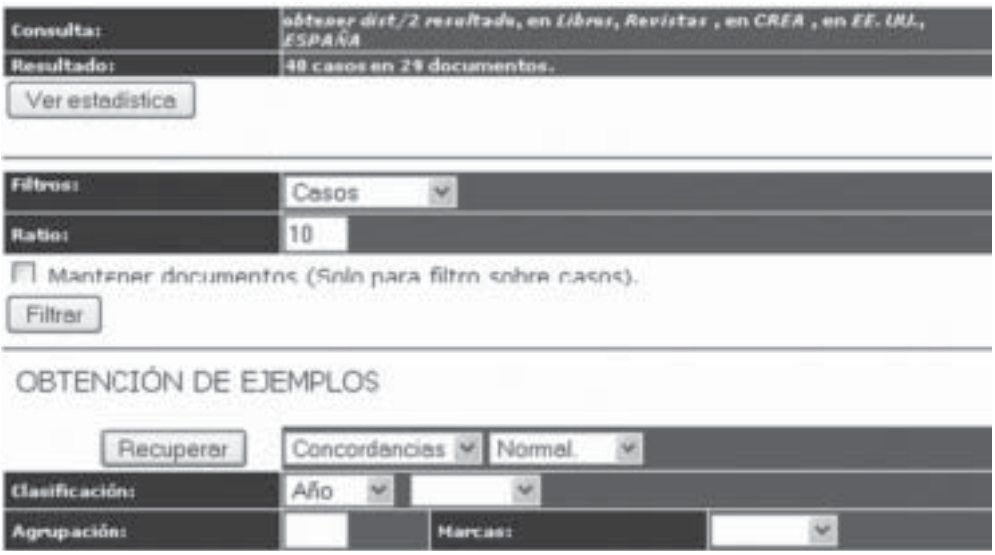
Tipos de consulta:

- por palabras o secuencias de palabras, p. ej. *bosque*, *tarjeta de crédito*
- con comodines (? y *), p. ej. *admirador**, *sobree*do*, *moz?*,
- por medio de operadores lógicos (Y, O, y NO), p. ej. *jugar Y ganar*, *jugar O perder, noche Y NO día*
- con el operador «dist/» que define la distancia máxima entre dos palabras en una sola línea de distancia, p. ej, *día dist/3 noche*

El sistema primero proporciona datos sobre el número de casos (ejemplos) y documentos encontrados, propone filtros y luego pasa a la recuperación de información concreta proveniente del corpus:

- documentos – i. e. datos bibliográficos,
- líneas de concordancias,
- párrafos, o
- agrupaciones – i. e. colocaciones, patrones más frecuentes, rasgos contextuales.

Se pueden ordenar las líneas de concordancias clasificándolas por datos bibliográficos o por el contexto en el que aparecen (de 5 a la izquierda hasta 5 a la derecha):



Cuadro 3: *Pantalla de resultados que proporciona datos sobre el número de documentos y casos encontrados en el CREA. La parte central sirve a los filtros y la parte de abajo para la recuperación de información concreta del corpus.*

Ng	CONCORDANCIA	AÑO
1	vidad productiva propuesta al fin y al resultado a obtener en el proceso de trabajo. En pues, una activi	** 1998
2	a el proceso de agregación seguido, con el fin de obtener, como resultado final, las cifras incluidas e	** 1992
3	hacerla cuando de ser cierta, la probabilidad de obtener el resultado que da el test fuese inferior a	** 1989
4	de obtener resultado positivo. b) Probabilidad de obtener resultado cero. c) Probabilidad de obtener 3	** 1998
5	resultados posibles y calcula: a) Probabilidad de obtener resultado positivo. b) Probabilidad de obtene	** 1998
6	el modo... 6g) Para alcanzar un objetivo: ... de obtener un resultado... 5g) Partiendo de una informac	** 2000
7	a de "Proposición dirigida a averiguar el modo de obtener un resultado cuando ciertos datos nos consid	** 2000
8	r varios implantes en un tiempo variable a fin de obtener un resultado definitivo. Todo este proceso se	** 1995
9	vejecido, una liposucción del mismo con el fin de obtener un resultado más perfecto. Y así por todas y	** 1995
10	i hubiera en él una tendencia, como si tratara de obtener un resultado, que sería el tipo ideal de anci	** 1979
11	suficientemente fuerte, tal vez se precipitaba en obtener el resultado. La respiración es la base, y lo	** 2001
12	e. Lo hizo más por resultar molesto que esperando obtener algún resultado de su búsqueda. Mi amada tani	** 1994
13	modificar el mapa de colores de la imagen, hasta obtener el resultado más óptimo. Figura 6.14 Otro asp	** 1994
14	al de poder, en un partido de masas, es necesario obtener un resultado tal que haga de nuevo, inapelabl	** 1993
15	tiempo y haberlos practicado repetidas veces para obtener algún resultado. Por esas razones, no los tra	** 1993
16	os con los que el resultor tiene que operar para obtener el resultado. Cuando intervienen más de dos s	** 2000
17	con marcas el renglón de "Datos" que se toma para obtener el resultado j). La última fila expresa el re	** 2000
18	bservó que la luz difusa era la más adecuada para obtener el resultado adecuado. En su caso de campo co	** 2000
19	cito del número de revelaciones del visinfín para obtener el resultado final de una vuelta cada 24 hora	** 1992
20	de medidas 3 (IMI) son aquellos que empleamos para obtener el resultado, una corrección. ESCALARES. En es	** 2000
21	o 3 (ESI o EFI) serán aquellos que empleamos para obtener el resultado, una corrección. PRODUCTO CARTESI	** 2000
22	cierta potencia. Muchas veces es suficiente para obtener este resultado el uso de un antidepresivo con	** 1998
23	de que la metódica aplicación de unas reglas para obtener un resultado. El juego es un camino de aprend	** 1987
24	operar con los valores contenidos en celdas, para obtener un resultado. Para introducir datos en una ho	** 1998
25	os fraccionamiento parece lo más conveniente para obtener un resultado soberanos capaz de resolver los	** 1994

Cuadro 4: *Visualización de líneas de concordancias de la consulta «obtener dist/2 resultado» en el CREA.*

Desde este programa es posible también obtener listados de frecuencias de palabras en el corpus: las primeras 1000, 5000, 10.000 palabras y un listado completo. Un fragmento del listado:

Orden		Frec. absoluta	Frec. normalizada
1.	**de**	9,999,518	65545.55
2.	**la**	6,277,560	41148.59
3.	**que**	4,681,839	30688.85
4.	**el**	4,569,652	29953.48
5.	**en**	4,234,281	27755.16

En el plazo 47 aparece la primera forma léxica – *años* (frecuencia absoluta: 203,027), seguida por:

58.	**vez**	163,538	1071.97
59.	**puede**	161,219	1056.76
64.	**parte**	148,750	975.03
65.	**tiene**	147,274	965.36
70.	**tiempo**	130,896	858.00

A la hora de consultar el CREA, es útil que el usuario tenga en cuenta las siguientes observaciones:

1) Existen limitaciones del sistema en cuanto a la recuperación de formas de alta frecuencia de aparición: el número máximo de documentos que puede suministrar una consulta es de 2000 y en cuanto al número de ejemplos, se proporcionan los primeros 1000 de cada consulta. Si el número de documentos excede a esta cifra, hay que recurrir a la página principal para restringir la consulta con varios criterios de selección o filtros disponibles.

2) En la versión de libre acceso no es posible hacer consultas por el lema de una palabra; así, no es posible obtener todas las formas de un lema concreto, por ejemplo del verbo *comer – como, comes, come* etc –. Tampoco hay excesivo margen en cuanto a la complejidad de las consultas. Es posible consultar el corpus por medio de operadores lógicos (Y, O, y NO) y comodines (? y *), pero el sistema rechaza las consultas (bien formadas) que proporcionarían un número más grande de casos.

3) Los desarrollos en curso, según el manual de consulta son: nuevos filtros estadísticos, recuperación sobre textos anotados con información lingüística (lema, clase de palabra, género, número etc.) y salvaguardia de los perfiles de consulta.

En la actualidad, Guillermo Rojo, uno de los protagonistas de la lingüística de corpus en España, coordina los trabajos de construcción del *Corpus del español del siglo XXI*, proyecto aprobado por la Asociación de Academias de la Lengua Española en su reunión en Medellín en marzo de 2007. Se planea que este corpus contenga textos de 2000 a 2011 con 300 millones de palabras. Entre otros, se ha revisado el criterio geográfico, pasando a los 30% para España y 70 % para la América Latina. De momento, esperemos que los mencionados desarrollos y otras mejoras posibilitadas hoy día por el avance de la informática, sean incorporados pronto al sistema existente para mejorar el acceso a la información en el CREA.

3.2 Corpus del español, de la Brigham Young University

El corpus ha sido construido por el profesor Mark Davies y contiene más de 100 millones de palabras procedentes de más de 20.000 textos del español que cubren el período comprendido entre los siglos XIII al XX. Del año 1200 a 1400, hay 20 millones de palabras; y para los siglos del 1500 al 1700 y 1800 al 1900 se han incluido 40 millones de palabras, respectivamente. Contiene textos literarios, hablados, de los periódicos y enciclopedias. Es de acceso libre y gratis en la página http://www.corpusdelespanol.org/ con sólo registrarse. Para los usuarios registrados, el sistema tiene unas funciones especiales: se guardan las consultas anteriores del usuario, puede añadir comentarios a las mismas para que los vean otros usuarios y puede comunicarse con otros usuarios, lo que resulta útil para su uso en las aulas.

La interfaz permite al usuario consultar el corpus de diferentes maneras:

- por palabras exactas o frases, p. ej. *bosque*, *polo norte*
- con comodines ? y *; p. ej. *averigu**
- por lemas, p. ej. *[subir]*
- por categoría gramatical, p. ej. *[vip*]*

Se pueden combinar los criterios (p. ej. *[nn*] fuerte*, *[vr*] * [pelota]*, *[vip*] hambre*), y hacer también las siguientes consultas:

- consultas de colocaciones,
- consultas por frecuencia y comparaciones de frecuencias de uso de palabras, frases y construcciones gramaticales,
- consultas basadas en la semántica y comparaciones de palabras diferentes y sinónimos.

Cuadro 5: *Consulta de la construcción «verbo en infinitivo + sustantivo pelota» o «[vr*] * [pelota]» en el Corpus del español*

3.3 Spanish Web Corpus integrado en la herramienta SketchEngine

SketchEngine es una herramienta de consulta de corpus desarrollada por la empresa Lexical Computing Ltd. (Kilgarriff 2004), disponible en http://www.sketchengine.co.uk/. El corpus de la lengua española dentro de la herramienta SketchEngine se llama *Spanish Web Corpus* y fue construido por Serge Sharoff de la universidad de Leeds. El corpus fue sintácticamente anotado y lematizado con el TreeTagger, mientras que el archivo de relaciones gramaticales fue preparado por Núria Bel y Hada Ross Salazar. La herramienta ofrece la posibilidad de una prueba de 30 días, mientras que las licencias académicas individuales cuestan GBP 55,25 por año.

Además de la visualización de las líneas de concordancias, la susodicha herramienta presta enorme ayuda en la investigación de contextos, colocaciones y comparaciones; el sistema ofrece asimismo listados de frecuencias y propone los ejemplos del corpus que tengan más relevancia (GDEX). En los cuadros de abajo se presentan dos de las funciones de la herramienta, el *WordSketch* y *Sketch-Diff*, respectivamente:

partido Spanish web corpus freq = 23916

object of	2324		subject of	1966	2.1	n_modifier	10406	2.2	modifies	1923	0.4
1.1											
☐ sacar	329	30.87	☐ haber	192	22.78	☐ comunista	1501	70.48	☐ se_puede	17	26.39
☐ tomar	257	36.34	☐ propugnará	5	21.5	☐ socialista	704	53.43	☐ compañero	46	26.27
☐ empatar	21	33.18	☐ luchar	14	19.82	☐ político	2084	51.12	☐ dirigente	26	22.06
☐ jugar	88	32.88	☐ estar	112	18.92	☐ popular	851	50.29	☐ bola	19	21.34
☐ ganar	88	29.09	☐ deber	69	18.1	☐ laborista	91	45.72	☐ líder	27	20.56
☐ fundar	32	25.86	☐ ser	280	17.55	☐ fútbol	224	42.71	☐ financiación	22	20.49
☐ disputar	16	23.72	☐ poder	99	16.8	☐ demócrata	125	41.23	☐ representante	30	20.16
☐ degollar	5	19.36	☐ tener-que	17	14.66	☐ liberal	210	38.66	☐ toma	29	19.85
☐ ver	81	18.25	☐ apoyar	12	12.48	☐ republicano	154	36.38	☐ concertación	10	19.79
☐ forjar	8	16.94	☐ aspirar	5	12.02	☐ oposición	190	35.01	☐ coalición	15	19.54
☐ gobernar	13	16.88	☐ era	24	11.44	☐ conservador	114	34.38	☐ sistema	92	18.57
☐ formar	38	16.07	☐ aceptar	12	10.58	☐ revolucionario	197	33.57	☐ despenalización	6	17.86
☐ organizar	19	15.45	☐ ganar	13	10.46	☐ izquierda	177	33.35	☐ legalización	8	17.75
☐ legalizar	6	15.2	☐ considerar	15	10.17	☐ bolchevique	47	31.82	☐ ley	53	16.66
☐ sentenciar	6	14.85	☐ tener	97	9.85	☐ mayoritario	62	29.84	☐ miembro	31	16.44
☐ perder	26	13.66	☐ defender	9	9.6	☐ políticos	28	28.42	☐ espíritu	29	16.18
			☐ llegar	17	9.6	☐ democrático	145	26.14	☐ resto	30	15.92

Cuadro 6: *Algunos de los contextos típicos del lema partido, clasificados por la frecuencia de su función gramatical y el cálculo estadístico salience, en la herramienta SketchEngine*

En el cuadro 6 se presentan algunas de las relaciones gramaticales más comunes (*object of*, *subject of*, *n_modifier*, *modifies*), con colocados y acceso directo a las concordancias relevantes.

object	7139	9011	3.7	4.6
objetivo	243	196	30.2	24.6
consenso	74	18	29.4	13.0
éxito	137	74	29.0	20.1
equilibrio	82	41	27.1	17.9
resultado	191	252	24.6	26.3
avance	36	28	25.5	12.9
victoria	57	62	25.0	24.6
acuerdo	115	39	24.5	9.9
dinero	7	124	2.2	23.5
meta	49	23	22.6	13.9
paz	84	56	21.1	15.5
empleo	18	76	8.5	20.1
progreso	50	17	19.8	9.1
premio	14	64	7.7	19.7
eficiencia	28	9	18.6	8.2
mejora	48	40	18.3	15.3
efecto	90	121	16.1	17.6
estabilidad	28	15	16.7	10.3
trabajo	19	185	0.4	16.5

Cuadro 7: *La función Sketch difference en la herramienta SketchEngine: comparación de lemas lograr y conseguir, listado de colocados en función de objeto (las columnas 2 y 4 para lograr, las columnas 3 y 5 para conseguir)*

El cuadro 7 presenta una comparación de colocados de los lemas *lograr* y *conseguir*, que son semánticamente bastante próximos; se puede observar el comportamiento de ciertos colocados respecto a sus rasgos combinatorios con los dos verbos. Así, por ejemplo, los colocados *objetivo, consenso, éxito, equilibrio* etc. se combinan más frecuentemente con *lograr* que con *conseguir*, mientras que *dinero, empleo, premio* y *trabajo* eligen más frecuentemente al verbo *conseguir.*

3.4 Otros corpus

Hay también una multitud de corpus que se han construido en el seno de las editoriales o grupos de investigación, y en general son de acceso restringido. Este apartado está destinado a presentar una selección de corpus que se pueden consultar.

3.4.1 Corpus Tècnic

Es un corpus de textos catalanes, ingleses y castellanos de diferentes ámbitos temáticos: informática, medio ambiente, derecho, medicina, genoma y economía. El corpus está anotado y lematizado; la parte española contiene 31.436.451 palabras. El programa de concordancias Bwananet permite al usuario consultar el corpus entero o definir los ámbitos y/o documentos que quiere consultar, pero la cantidad de concordancias que se pueden recuperar en acceso abierto es de 50 casos. Se desarrolló en la Universitat Pompeu Fabra (http://www.iula.upf.edu/corpus/corpusuk.htm) y está disponible en http://bwananet.iula.upf.edu/indexen.htm.

3.4.2 COLA

http://colam.org/transkripsjoner-espannol.html

El fin principal del proyecto COLA (*Corpus Oral de Lenguaje Adolescente*) es recoger el habla de jóvenes madrileños comprendidos entre los 13 y 19 años, así como de algunas capitales latinoamericanas (Buenos Aires, Santiago de Chile, Guatemala, La Habana), para construir un corpus del habla juvenil para la investigación, accesible en Internet. Los usuarios tienen que registrarse y, de momento, es posible hacer consultas en el corpus de Madrid (COLAm), que contiene alrededor de 400.000 palabras transcritas. También son accesibles secuencias de grabaciones y un listado de frecuencias. Un ejemplo de transcripción del contexto del lema *nena* en el corpus COLA:

<p MALCE2J02-> <Comment desc=«voces de fondo«/> hala qué fuerte mira habla se escucha todo lo que lo que se dicen a kilómetros **nena** <p MALCE2J01-> madre mía qué guapo <p MALCE2J02-> mira <p MALCE2G01-> el tuto <p MALCE2G01> si no sé qué venga y tal habla para jugar <p MALCE2G03->

3.4.3 ARTHUS

El corpus Arthus (*Archivo de Textos Hispánicos de la Universidad de Santiago de Compostela*, http://www.bds.usc.es/corpus.html) se ha construido en la Universidad de Santiago de Compostela para la compilación de la *Base de datos sintáctios* (BDS). Contiene 1.450.000 palabras en los siguientes porcentajes: prosa (37 %), ensayo (18 %), dramática (14 %), periódicos (11 %) y la componente hablada (19 %). Un 79 % de textos son de procedencia española y 21 % de procedencia latinoamericana. Es posible consultarlo entrando en la base de datos BDS (http://www.bds.usc.es/bds.html) o ADESSE (http://adesse.uvigo.es), que es la continuación del proyecto anterior.

3.4.4 El Corpus LexEsp

El corpus LexEsp (*Léxico informatizado del español*) se ha desarrollado en colaboración entre la Universitat de Barcelona y la Universitat Politècnica de Catalunya. Contiene textos de varios géneros literarios, noticias, prensa y artículos científicos. Consta de más de 5,5 millones de palabras del español contemporáneo. Se ha publicado en formato electrónico (Sebastián et al. 2000), pero también es posible consultar el contenido del corpus en http://www.lsi.upc.es/~nlp/tools/corpus-es.php. Por otra parte, una versión parcial del corpus en formato textual está disponible en

http://www.psico.uniovi.es/Dpto_Psicologia/metodos/soft/corpus/base/.

3.4.5 Corpus Trilingüe Paralelo GRIAL y SenSem (Corpus del español anotado sintácticamente y semánticamente)

http://grial.uab.es/recursos.php?idioma=es

http://grial.uab.es/fproj.php?id=1&idioma=es

En la Universidad Autónoma de Barcelona, se ha desarrollado dos corpus:

- GRIAL, que es un corpus paralelo para el inglés, el español y el catalán; comprende 2.257.498 palabras para los tres idiomas y está anotado automáticamente a nivel morfosintáctico;
- SenSem: corpus que incluye textos del ámbito periodístico (*El Periódico de Cataluña*). De este corpus se han seleccionado 25.000 frases para hacer el análisis de los 250 verbos más frecuentes del español actual.

3.4.6 CODICACH

El Corpus Dinámico del Castellano de Chile (CODICACH) se desarrolla en la Universidad de Concepción en Chile desde 1997 en adelante. Es un corpus sincrónico del español escrito de Chile, compuesto de cerca de 800 millones de palabras. La mayor parte de los textos fueron escritos entre 1997 y 2003. El corpus contiene textos de fuentes escritas y una parte oral transcrita. Se ha hecho un gran esfuerzo en eliminar en la medida de lo posible textos de autores no chilenos. El corpus está compuesto de archivos de texto plano y se planifica una transformación al formato xml y la incorporación de la metainformación. Para más información, ver http://www2.udec.cl/~ssadowsky/codicach.html y para acceder al corpus hay que ponerse en contacto con el autor.

4. El futuro de corpus y el reto de Internet

En los últimos años, sin embargo, la lingüística de corpus se ocupa cada vez más de la exploración de datos ofrecidos por la fuente global, la red. En líneas generales, los expertos en lingüística de corpus y lingüística computacional están de acuerdo en que los corpus del futuro deberían ser:

- más grandes y mejores
- provenientes de los datos de Internet
- de dos tipos prevalecientes: *abierto/monitor* o *ad hoc*

El tamaño de los corpus de referencia hoy en día alcanza ya cifras vertiginosas, que en algunos casos sobrepasan ya a mil millones de formas (por ejemplo ukWac, deWac, itWac, incorporados en SketchEngine). Como es lógico, tal cantidad de datos requiere un continuo desarrollo paralelo de las herramientas de consulta, y análisis estadísticos de la información que tengan en cuenta la cantidad de datos.

Internet, o la red, es hoy una fuente enorme de materiales lingüísticos. Aunque existen argumentos pro y contra el uso de los textos de la red, y el diseño de un corpus también depende de las finalidades de una investigación concreta, se reconoce que los datos de la red son útiles; además, los derechos de autor, que son un tema de crucial importancia en la construcción de corpus textuales «tradicionales», tienen un aspecto bastante diferente en Internet.

Por la cantidad de materiales hoy disponibles y el estado de las tecnologías, el desarrollo y la construcción de corpus nuevos se mueve en dos líneas generales: *corpus abiertos* y *corpus construidos ad hoc*. Los primeros tienen un diseño que permite que los materiales entren y salgan del corpus dependiendo de la fecha de su creación u otro crite-

rio tangible, mientras se mantiene la representatividad diseñada del corpus. Los segundos explotan otro aspecto del desarrollo informático: la posibilidad de crear uno mismo su propio corpus de un modo rápido. La investigación actual sobre el uso de Internet para la construcción gira, sobre todo, en torno a los talleres anuales Web as Corpus, que se celebran desde 2005.

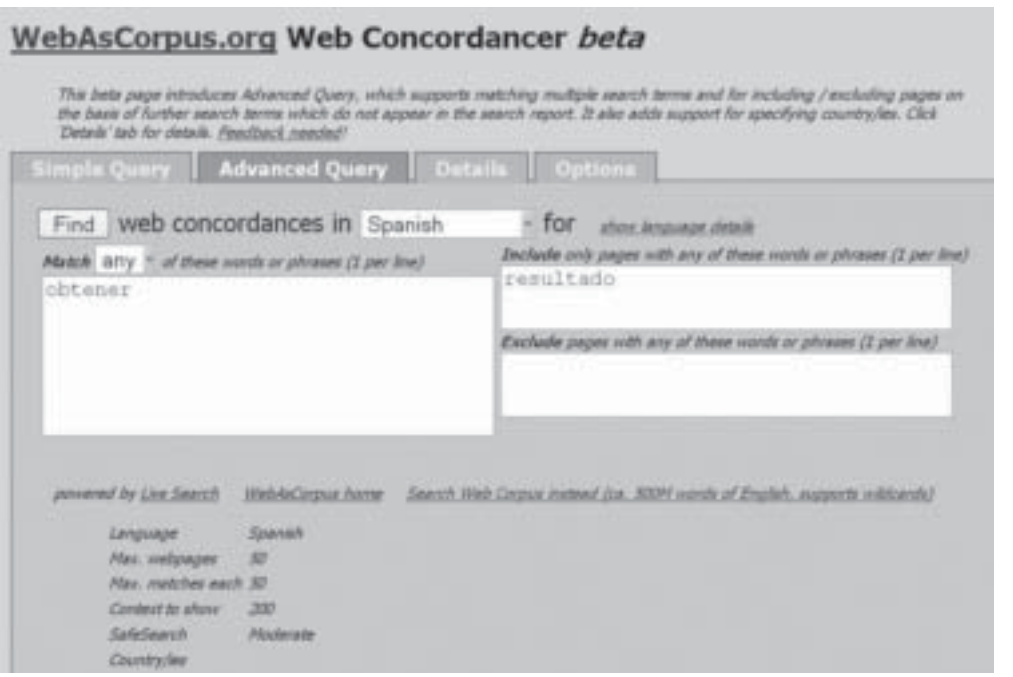
A continuación, pasamos a revisar brevemente una selección de los desarrollos arriba descritos.

4.1 Web Concordancer beta

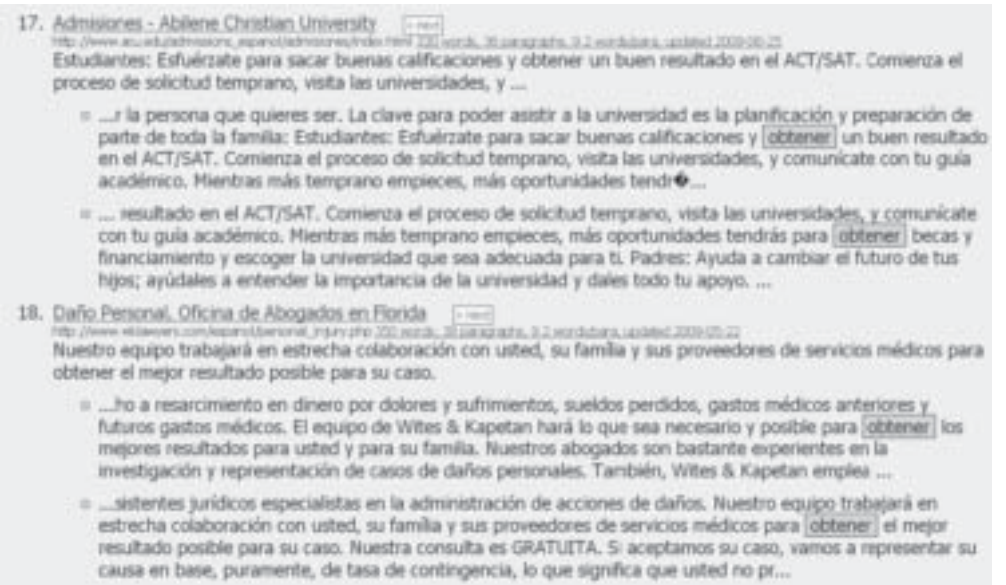
http://webascorpus.org/searchwac.html

http://webascorpus.org/

Herramienta de consulta de Internet con visualización de concordancias. De momento, se puede hacer consultas en 34 lenguas. El programa apoya consultas de varias palabras a la vez y tiene filtros de país. Es posible también descargar los resultados en formato textual. Adelante, se presentan la página de consulta y el resultado.



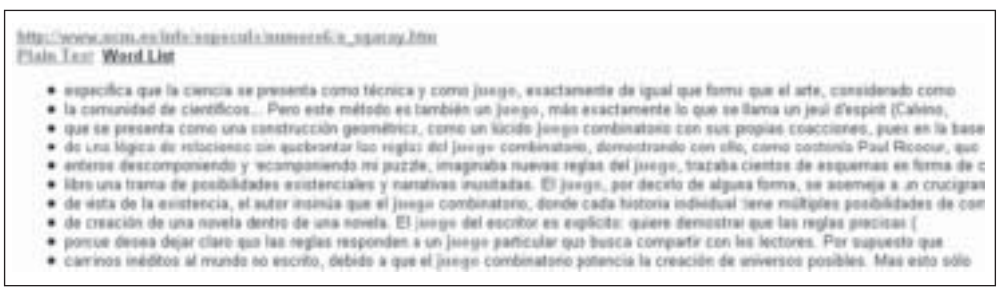
Cuadro 8: *La consulta obtener y resultado en la herramienta Web Concordancer beta*



Cuadro 9: *Visualización de las líneas de concordancias en Web Concordancer beta*

4.2 WebCorp

WebCorp es una herramienta desarrollada por Birmingham City University y sirve para consultar Internet como corpus (Renouf et al. 2007). Para manejar las consultas hay varias opciones: comodines, grupos de letras o palabras, filtros, nombre de dominio, tipo de texto, buscador. Resultado típico de una consulta simple:



Cuadro 10: Consulta *juego* en Webcorp, páginas españolas, contexto 10 palabras.

4.3 WebBootCaT

WebBootCaT es la versión web de la herramienta BootCaT (*Bootstrap Corpora and Terms from the Web*) que permite al usuario crear su propio corpus *ad hoc*. Se ha creado sobre todo para los traductores que a menudo tienen que recurrir a Internet para solucio-

nar las preguntas que los diccionarios generales no suelen resolver (Baroni et al. 2006a, Baroni et al. 2006b).

El proceso de construcción del corpus es el siguiente:

1) El usuario define las palabras claves o *seed words* (literalmente «palabras semilla») (Cuadro 11).
2) Se recuperan las páginas web.
3) Se recuperan textos de las páginas seleccionadas (Cuadro 12).
4) TreeTagger: anotación y lematización del corpus (por ahora, TreeTagger existe para el análisis de varias lenguas; para más información, v. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/).

En la página principal de construcción del corpus, el usuario define las palabras clave, el idioma, la herramienta de anotación del corpus y el nombre del corpus:

Cuadro 11: *Página principal de la entrada de datos en la herramienta WebBootCaT*

Corpus built

Your corpus was built successfully.

Corpus name	Prueba_DEPORTE_01
Size	599 kB
Word count	117338
Web pages retrieved	53
Build time	02:29
Access URL	http://beta.sketchengine.co.uk/auth/corpora/run.cgi/first_form?corpname=wbc/amebis00/EqoboQZz

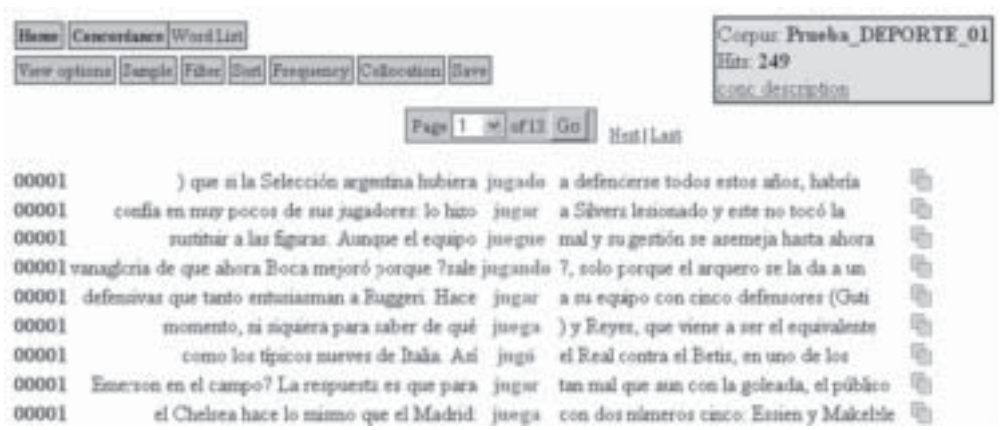
Download the corpus in raw format

Download the corpus in vertical format

Extract keywords

Cuadro 12: *Información sobre el corpus creado ad hoc a base de palabras clave en la herramienta WebBootCaT*

Como herramienta WebBootCaT está integrada en la herramienta SketchEngine (www.sketchengine.co.uk), es posible consultar las concordancias e incluso utilizar las siguientes funciones: Concordancias (filtro, clasificación, frecuencias, colocaciones), Word List –listados de palabras, lemas o marcas–, Find X, Extract Keyterms & la posibilidad de crear un segundo corpus de los términos seleccionados enseguida. También es posible descargar el corpus entero y consultarlo con otros programas de concordancias.



Home | Concordance | Word List

View options | Sample | Filter | Sort | Frequency | Collocation | Save

Corpus: **Prueba_DEPORTE_01**
Hits: **249**
conc description

Page 1 of 13 Go | first | last

00001	) que si la Selección argentina hubiera	jugado	a defenderse todos estos años, habría
00001	confía en muy pocos de sus jugadores: lo hizo	jugar	a Silveri lesionado y este no tocó la
00001	suntuar a las figuras. Aunque el equipo	juegue	mal y su gestión se asemeja hasta ahora
00001	vanagloria de que ahora Boca mejoró porque ?sale	jugando	?, solo porque el arquero se la da a un
00001	defensivas que tanto entusiasman a Ruggeri. Hace	jugar	a su equipo con cinco defensores (Guti
00001	momento, ni siquiera para saber de qué	juega	) y Reyes, que viene a ser el equivalente
00001	como los típicos nuevos de Italia. Así	jugó	el Real contra el Betis, en uno de los
00001	Emerson en el campo? La respuesta es que para	jugar	tan mal que aun con la goleada, el público
00001	el Chelsea hace lo mismo que el Madrid:	juega	con dos números cinco: Essien y Makelele

Cuadro 13: *Parte de la pantalla con la consulta del corpus, construido en la herramienta WebBootCaT; visualización en la herramienta SketchEngine*

Cuadro 14: *Las palabras clave clasificadas por frecuencia, resultado del corpus construido ad hoc en la herramienta WebBootCaT*

5. Conclusión

A lo largo del presente artículo, se han presentado varios recursos lingüísticos que están a disposición de los usuarios de la lengua española en Internet. En el primer bloque, tratamos los diccionarios en formato electrónico, y en el segundo, los corpus, que permiten al usuario observar las palabras consultadas en sus contextos. Entre los más divulgados, están el CREA de la RAE, el Corpus del español y Spanish Web Corpus integrado en la herramienta SketchEngine. Se muestran varias funciones de las herramientas de consulta de corpus, junto a la visualización de los resultados. El tercer apartado está dedicado a la presentación de los recursos que van más allá de los diccionarios y corpus tradicionales –estos utilizan Internet como su fuente principal de textos–. Se presentan igualmente algunas herramientas que facilitan la consulta de Internet (Web Concordancer beta, WebCorp y WebBootCaT).

BIBLIOGRAFÍA

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R., Almela, M. (2005): *Frecuencias del español: Diccionario y estudios léxicos y morfológicos.* Madrid: Editorial Universitas.

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006a): «WebBootCaT: instant domain-specific corpora to support human translators». En: *Proceedings of EAMT 2006*, Oslo, 247–252.

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006b): «WebBootCaT: a web tool for instant corpora». En: *Proceedings / XII Euralex International Congress,* Alessandria: Edizioni dell'Orso, 123–131.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. (2004): «The Sketch Engine» En: *Proceedings / XI Euralex International Congress,* Lorient: Université de Bretagne-Sud, 105–116.

Lavid, J. (2005): *Lenguaje y nuevas tecnologías: Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Madrid: Cátedra.

Renouf, A., A. Kehoe & J. Banerjee (2007): «WebCorp: an integrated system for web text search» En: C. Nesselhauf, M. Hundt & C. Biewer (eds.), *Corpus Linguistics and the Web.* Amsterdam: Rodopi, 47–67.

Sebastián, N., Cuetos, F., Martí, M. A., Carreiras, M. F. (2000): *LEXESP: Léxico informatizado del español*. Edición en CD-ROM. Barcelona: Edicions de la Universitat de Barcelona.

Enlaces a los corpus y herramientas descritas en el artículo

[fecha de consulta: 15 de junio de 2009]:

Arthus (*Archivo de Textos Hispánicos de la Universidad de Santiago de Compostela*):

Corpus Oral de Lenguaje Adolescente (COLA): http://colam.org/transkripsjoner-espannol.html

Corpus Tècnic del IULA de la UPF (CT-IULA), datos obtenidos a través de Bwananet en el período junio/2009: http://bwananet.iula.upf.edu/indexen.htm

Corpus Trilingüe Paralelo GRIAL: http://grial.uab.es/recursos.php?idioma=es

Davies, M. (n. d.): *Corpus del Español*. (Brigham Young University) En: Corpus del español: http://www.corpusdelespanol.org/

LexEsp corpus: http://www.lsi.upc.es/~nlp/tools/corpus-es.php

Real Academia Española: *Banco de datos (CREA)* [en línea]. *Corpus de referencia del español actual.* <http://www.rae.es>

SenSem (*Corpus del español anotado sintácticamente y semánticamente*):

http://grial.uab.es/fproj.php?id=1&idioma=es

Lexicom Lexical Computing (n. d.): *SketchEngine*. En: *SketchEngine*: http://www.sketchengine.co.uk/

Web as Corpus: http://webascorpus.org/

WebCorp: http://www.webcorp.org.uk/

INTERNET IN JEZIKOVNA SREDSTVA V ŠPANSKEM JEZIKU: SLOVARJI IN KORPUSI

Ključne besede: jezikovna sredstva, slovar, referenčni korpus, internet, španski jezik

Referenčna dela in različna jezikovna sredstva (besedišče, slovnice, zbirke besedil, korpusi) predstavljajo povezavo med jezikom in uporabnikom, ki išče podatke o maternem ali tujem jeziku. Nove tehnologije odpirajo vrata v nov svet in omogočajo preučevanje jezika z drugačnih, alternativnih vidikov ter predstavljajo dejavnik, ki ga je potrebno upoštevati pri širjenju jezikovne informacije. Pričujoči članek se ukvarja z opredelitvijo in opisom izbranih jezikovnih sredstev na področju španskega jezika, namenjenih uporabnikom, ki iščejo podatke o španščini za katerokoli jezikovno dejavnost. Članek opisuje in opredeljuje tista sredstva, ki so brezplačno ali za primerno ceno dostopna na spletu celotni jezikovni skupnosti. V prvem delu so predstavljeni elektronski slovarji, v drugem delu pa korpusi, ki uporabnikom omogočajo opazovanje izbranih besed v kontekstih. Avtorica predstavi najbolj znane korpuse španskega jezika, kot so CREA Španske kraljeve akademije, korpusa *Corpus del español* in *Spanish Web Corpus*, vključen v orodje *SketchEngine*, in nekatere funkcije orodij za uporabo korpusov. V tretjem delu avtorica razmišlja o novih smereh razvoja v prihodnosti in predstavi nekatera orodja, ki omogočajo boljšo in enostavnejšo uporabo Interneta (*Web Concordancer beta* in *WebCorp y WebBootCaT*).